

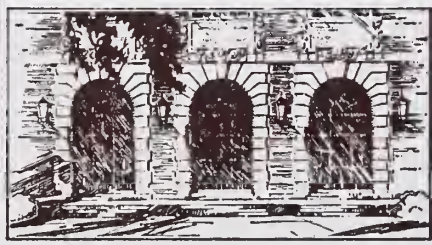
LIBRARY OF THE
UNIVERSITY OF ILLINOIS
AT URBANA-CHAMPAIGN

510.84

I l 6 r

no.100-110

cop.3





Digitized by the Internet Archive
in 2013

<http://archive.org/details/roundofferroracc103greg>

34
r
3
5

UNIVERSITY OF ILLINOIS
GRADUATE COLLEGE
DIGITAL COMPUTER LABORATORY

REPORT NO. 103

ROUND-OFF-ERROR ACCUMULATION IN ITERATIVE PROCEDURES

by

Robert T. Gregory and A. H. Taub

October 31, 1960

This work was supported in part by
the National Science Foundation
under Grant G-9503.

ROUND-OFF-ERROR ACCUMULATION IN ITERATIVE PROCEDURES

Robert T. Gregory and A. H. Taub

I. Introduction. In an unpublished paper (see appendix) Taub and Wilf investigate stopping procedures for linear iterative processes. In particular, they study iterative methods for solving a system of linear algebraic equations numerically, where each method produces a sequence of vectors $\{u_n\}$ with the property that the solution vector is approximated by the limit u_∞ of the sequence. They point out that unless the stopping criterion is chosen carefully, the number of iterations performed may be much larger than necessary due to misleading information given by the cancellation of roundoff error.

The mathematical model chosen for their investigation is extremely simple (see appendix, section II). It is stated that the single equation

(1)
$$u_{n+1} = bu_n,$$

where

(2)
$$1/2 < b < 1,$$

is perfectly general if b is thought of as the largest eigenvalue of the iteration operator b . In this case

(3)
$$u_\infty = 0.$$

Machine calculation of (1) actually produces

(4)
$$\bar{u}_{n+1} = b\bar{u}_n + \delta_n,$$

where δ_n is the roundoff error. Under the hypotheses that the δ_n are uncorrelated random variables with common probability density function $f(\delta)$, mean value zero, common variance σ^2 , and uniform distribution, it is shown (see appendix, sections II and III) that stopping procedure P_α (defined below) allows more iterations, in general, than stopping procedure P_β (also defined below). These extra iterations, of course, are wasted.

One purpose of the present investigation is to attempt to obtain machine results which verify the theoretical result of the previous paragraph and thus validate the assumptions made about the δ_n . However, results described below indicate that, for the model chosen, the δ_n are not uniformly distributed random variables when n is large and the iterations approach convergence.

II. The Machine Program. In order to study the mathematical model mentioned above, a program was written for the ILLIAC to carry out the computation indicated in (4). It is possible, for a given value of b , to try many cases, each one with a different starting value u_0 (generated as a pseudo random number) where b satisfies (2). In each case, for $n = 1, 2, \dots, N$, the quantities $|\bar{u}_{n+1} - \bar{u}_n|$ and $|\bar{u}_{n+1} - \bar{u}_\infty| = |\bar{u}_{n+1}|$ are computed.

It is shown in (9) of the appendix that the variance of $\bar{u}_{n+1} - \bar{u}_n$ is less than the variance of \bar{u}_{n+1} . Hence, noting (3), a stopping criterion

(P α): Stop the iteration when $\bar{u}_{n+1} - \bar{u}_n$ is "pure noise,"

should allow overiteration as compared with a stopping criterion

(P β): Stop the iteration when $\bar{u}_{n+1} - u_\infty = \bar{u}_{n+1}$ is "pure noise." *

The difference in the number of iterations using the two criteria is given by (14) of the appendix.

It was anticipated that the quantities $|\bar{u}_{n+1} - \bar{u}_n|$ and $|\bar{u}_{n+1}|$ would remain monotone decreasing until they became small and roundoff-error accumulation became significant, at which time fluctuations would occur. These fluctuations, in each case, would be detected and used as a criterion for stopping the iteration. The theory states that the fluctuation in $|\bar{u}_{n+1} - \bar{u}_n|$ should begin many iterations later than the fluctuation in $|\bar{u}_{n+1}|$.

* See equations (10) and (12) of the appendix for the definition used for the value of n at which $\bar{u}_{n+1} - \bar{u}_n$ and \bar{u}_{n+1} are "pure noise."

Because of this fact, a third quantity was computed by the program, namely $|\bar{u}_{n+1} - \bar{u}_n + \Delta_n|$, where Δ_n is a uniformly distributed random variable generated so as to make the variance of $|\bar{u}_{n+1} - \bar{u}_n + \Delta_n|$ equal the variance of $|\bar{u}_{n+1}|$, that is, so as to make the fluctuation in $|\bar{u}_{n+1} - \bar{u}_n + \Delta_n|$ begin at the same time as the fluctuation in $|\bar{u}_{n+1}|$, on the average. In other words, it was hoped that by introducing some random noise of the form Δ_n one might make up for the cancellation of roundoff noise due to the subtraction performed in forming $|\bar{u}_{n+1} - \bar{u}_n|$. Thus, a stopping criterion based on $|\bar{u}_{n+1} - \bar{u}_n + \Delta_n|$ would allow fewer iterations, on the average, than the widely-used criterion based on $|\bar{u}_{n+1} - \bar{u}_n|$.

III. Numerical Results. The machine actually normalized the quantities mentioned in the previous section by dividing them by $|u_0|$. Since the results of the first few iterations were of little interest, the printing portion of the main loop was by-passed except when any one of the three inequalities,

$$(5) \quad \left| \frac{\bar{u}_{n+2}}{u_0} \right| - \left| \frac{\bar{u}_{n+1}}{u_0} \right| \geq 0$$

$$(6) \quad \left| \frac{\bar{u}_{n+2} - \bar{u}_{n+1}}{u_0} \right| - \left| \frac{\bar{u}_{n+1} - \bar{u}_n}{u_0} \right| \geq 0,$$

or

$$(7) \quad \left| \frac{\bar{u}_{n+2} - \bar{u}_{n+1} + \Delta_{n+1}}{u_0} \right| - \left| \frac{\bar{u}_{n+1} - \bar{u}_n + \Delta_n}{u_0} \right| \geq 0,$$

was satisfied.

Table I shows a typical set of results, where $b = 0.9$ and $u_0 = 0.0966\ 2045\ 7638$. The numbers printed should be multiplied by 2^{-39} , since the ILLIAC word contains a sign and thirty-nine binary digits, and the results were printed as integers.

It is seen that \bar{u}_{n+1} is monotone decreasing until it reaches a value such that rounded multiplication by b produces the same value again. When this occurs, the differences become zero, and the pattern of behavior expected for $|\bar{u}_{n+1} - \bar{u}_n|$ does not appear. More than seventy-five machine runs were made with various values of b and u_0 , and the results obtained in each case agree with the case shown.

Figure 1 shows the behavior of \bar{u}_{n+1} graphically. Here we have results for $b = 0.9375$ and $u_0 = 0.0839\ 0603\ 8013$. Notice that after 330 iterations the differences are constant and equal to 2^{-39} . At this point it is predictable what \bar{u}_{n+1} will be for a given \bar{u}_n . Ultimately, the constant value $\bar{u}_n = 8(2^{-39})$ is reached. Since the ILLIAC rounds by adding 2^{-40} to the product and then truncating, we see that

$$\begin{aligned}
 (8) \qquad (0.9375)(8)(2^{-39}) + 2^{-40} &= \left(\frac{15}{16}\right)(2^{-36}) + 2^{-40}, \\
 &= 15(2^{-40}) + 2^{-40}, \\
 &= 8(2^{-39}),
 \end{aligned}$$

and $\bar{u}_{n+1} = \bar{u}_n$ from this point on.

IV. An Analysis of the Difficulty. It was expected that when \bar{u}_{n+1} became "pure noise" we would observe fluctuations in its value rather than a leveling-off at a constant value. It is this leveling-off that explains our failure to observe the expected behavior of $|\bar{u}_{n+1} - \bar{u}_n|$.

One might ask whether or not it was accidental that \bar{u}_{n+1} , in the example of Figure 1, reached the critical value $8(2^{-39})$, and if so, could this be avoided.

TABLE I

The numbers in the last three columns should be multiplied by 2^{-39} .

n	$\left \frac{\bar{u}_{n+1}}{u_0} \right $	$\left \frac{\bar{u}_{n+1} - \bar{u}_n}{u_0} \right $	$\left \frac{\bar{u}_{n+1} - \bar{u}_n + \Delta_n}{u_0} \right $
190	1002	113	113
191	904	103	113
193	726	83	83
194	654	73	93
195	590	63	63
196	526	63	73
197	478	51	51
198	426	51	63
199	382	41	41
200	342	41	51
201	310	31	41
202	278	31	31
203	250	31	41
204	226	21	31
205	206	21	31
206	186	21	21
207	166	21	31
208	146	21	31
209	134	11	11
210	126	11	11
211	114	11	21
212	102	11	11
213	94	11	11
214	82	11	21
215	74	11	21
216	62	11	21
217	50	11	21
218	42	11	11
219	42	0	0
220	42	0	0
221	42	0	11

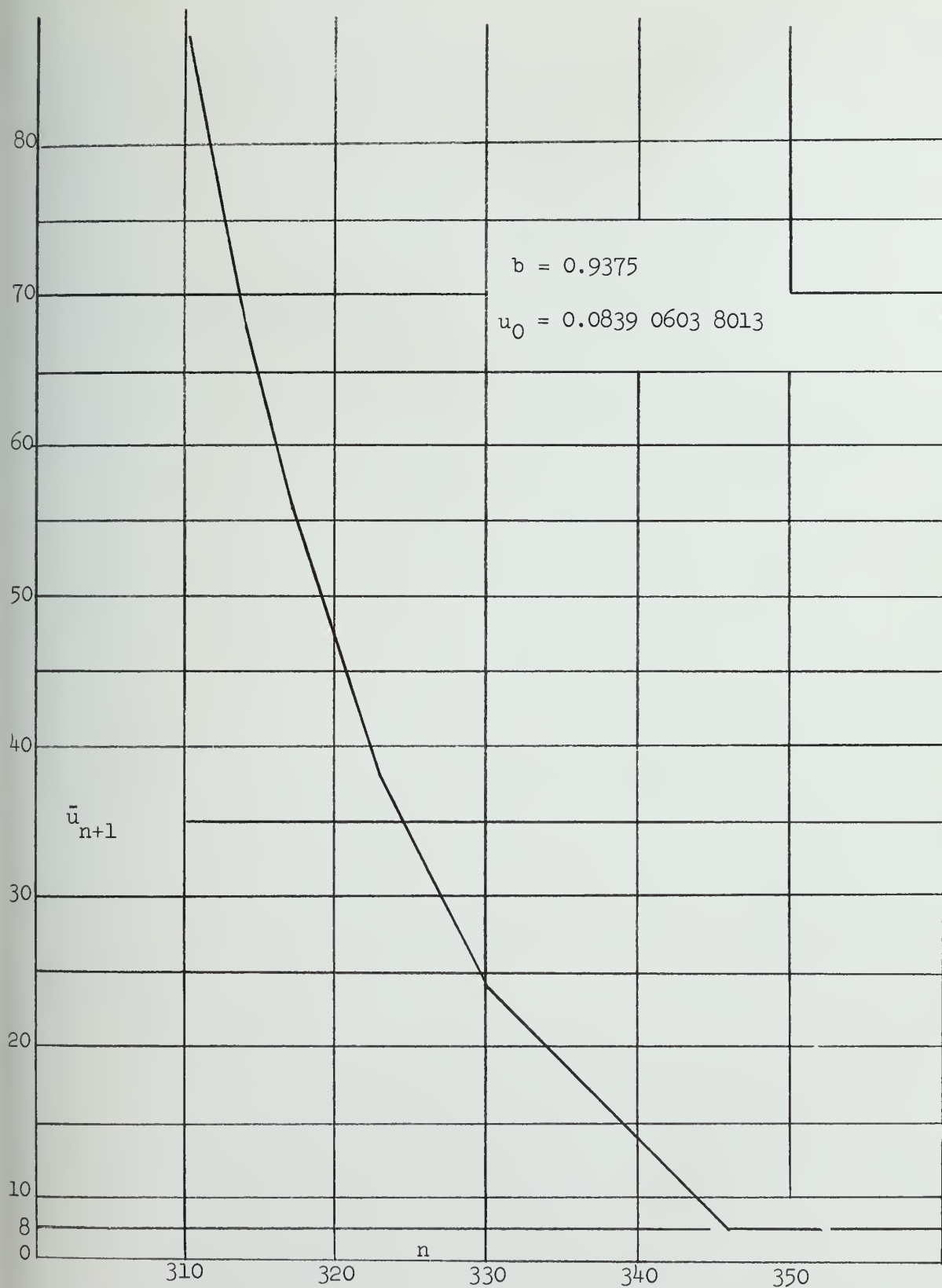


Figure 1

n is the number of iterations. \bar{u}_{n+1} is the approximation to the solution and should be multiplied by 2^{-39} .

The answer appears to be negative, because after 330 iterations

$$(9) \quad \bar{u}_n = 24(2^{-39}),$$

and beginning at this point,

$$(10) \quad \bar{u}_{n+1} = \bar{u}_n - 2^{-39},$$

so that sixteen iterations later

$$(11) \quad \bar{u}_n = 8(2^{-39}),$$

as could be predicted. This behavior was observed in every case tried.

Let us examine this analytically. Set

$$(12) \quad \bar{u}_n = \sum_{i=1}^{39} \alpha_i 2^{-i},$$

and consider the case

$$(13) \quad \begin{aligned} b &= 0.9375 \\ &= 1 - 2^{-4}. \end{aligned}$$

Then

$$(14) \quad \begin{aligned} \bar{u}_{n+1} &= \left[(1 - 2^{-4}) \sum_{i=1}^{39} \alpha_i 2^{-i} + 2^{-40} \right] \text{ truncated} \\ &= \left[\bar{u}_n + 2^{-40} - \sum_{i=1}^{39} \alpha_i 2^{-i-4} \right] \text{ truncated} \end{aligned}$$

In order to see what the truncated value becomes, we write the binary digits in the following manner:

$$\begin{array}{cccccccccccc|cccc} \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 & \alpha_5 & \alpha_6 & \dots & \alpha_{37} & \alpha_{38} & \alpha_{39} & & 1 & 0 & 0 & 0 \\ & & & & & & & & & & & \alpha_{36} & \alpha_{37} & \alpha_{38} & \alpha_{39} \end{array}$$

The truncation takes place to the right of the thirty-ninth digit, following the subtraction.

There are three cases:

Case I $\alpha_{36} = 0$

Case II $\alpha_{36} = 1, \alpha_{37} = \alpha_{38} = \alpha_{39} = 0$

Case III $\alpha_{36} = 1$, at least one of α_{37}, α_{38} , or α_{39} does not vanish.

In Cases I and II the truncated value becomes

$$(15) \quad \bar{u}_{n+1} = \bar{u}_n - \sum_{i=1}^{35} \alpha_i 2^{-i-4}.$$

In Case III it becomes

$$(16) \quad \bar{u}_{n+1} = \bar{u}_n - \sum_{i=1}^{35} \alpha_i 2^{-i-4} - 2^{-39}.$$

Thus (10) will be satisfied if

$$(17a) \quad \bar{u}_n = 0.000 \dots 010xxx,$$

$$(17b) \quad \bar{u}_n = 0.000 \dots 011000,$$

or

$$(17c) \quad \bar{u}_n = 0.000 \dots 001xxx,$$

that is to say, whenever

$$(18) \quad 8(2^{-39}) < \bar{u}_n \leq 24(2^{-39}).$$

Once this situation occurs, it is certain that the critical value $8(2^{-39})$ will be reached during some subsequent iteration.

V. Conclusion. It appears from these results that for the model chosen the assumption that the δ_n are uncorrelated, uniformly-distributed random variables is invalid. It may be that for a more complex model (possibly one consisting of a system of two equations) the assumptions made about roundoff errors are valid. This will be examined in a subsequent report.

APPENDIX

STOPPING PROCEDURES FOR LINEAR ITERATIVE PROCESSES

H. S. Wilf and A. H. Taub

Introduction

In recent years, rather considerable attention has been paid to developing more rapidly convergent algorithms for solving the system

$$(1) \quad x = Ax + b$$

numerically, where x, b are column vectors, and A is an $N \times N$ matrix. These methods all have for their object the construction of an iterative scheme,

$$(2) \quad u_{n+1} = Bu_n + C,$$

where the largest eigenvalue of B is as much less than unity in magnitude as possible, and such that the solution x , of (1) can be obtained by inspection from the limit u_{∞} of (2).

The purpose of this note is to point out that unless the stopping criterion for (2) is carefully chosen, the gain obtained from reduction of the largest eigenvalue will be entirely thrown away because of misleading information given by the cancellation of roundoff error.

Statement of the Problem

In (2), let us take $C = 0$, so that $u_{\infty} = 0$. This will not affect the final results. Next, let us treat the case $N = 1$, so that B is a scalar b . While seemingly drastic, this is actually perfectly general if b is thought of as the largest eigenvalue of B , as can be seen by diagonalizing B in (2).

The iteration (2) then generates the numbers

$$(3) \quad u_n = b^n u_0.$$

What is actually calculated, however, is

$$(4) \quad \bar{u}_{n+1} = b\bar{u}_n + \delta_n,$$

where δ_n is roundoff error, for which we assume

H1: The δ_n have a common distribution function $f(\delta)$ with finite variance σ^2 .

H2: The δ_n are uncorrelated random variables.

Consider the following two stopping procedures:

(P α): Stop the iteration when $\bar{u}_{n+1} - \bar{u}_n$ is "pure noise."

(P β): Stop the iteration when $\bar{u}_n - \bar{u}_\infty = \bar{u}_n$ is "pure noise."

We see that P α is the one normally used because the limit u_∞ is unknown. We will now show that P α will always allow the calculation to go on much longer than P β would, the extra iterations, of course, being quite wasted since the solution is not being improved. This phenomenon, as will be seen below, is directly the result of "cancellation" of noise in $\bar{u}_{n+1} - \bar{u}_n$ by the subtraction involved, which gives the iterates a deceptively significant appearance.

I. Analysis of P α and P β

The solution of (4) is

$$(5) \quad \bar{u}_n = b^n u_0 + \left[b^{n-1} \delta_0 + b^{n-2} \delta_1 + \dots + b \delta_{n-2} + \delta_{n-1} \right].$$

Thus,

$$(6) \quad \bar{u}_{n+1} - \bar{u}_n = b^n (b-1) u_0 + \left[b^{n-1} (b-1) \delta_0 + \dots + b (b-1) \delta_{n-2} + (b-1) \delta_{n-1} + \delta_n \right],$$

and the variance is (since the δ_n are uncorrelated)

$$\begin{aligned}
 (7) \quad \sigma_n^2(\alpha) &= (b-1)^2 \sigma^2 \left\{ \frac{1-b^{2n}}{1-b^2} \right\} + \sigma^2, \\
 &= \frac{\sigma^2}{1+b} \left\{ 2 - b^{2n} + b^{2n+1} \right\}.
 \end{aligned}$$

For process $P\beta$, we read directly from (5) the noise level

$$(8) \quad \sigma_n^2(\beta) = \sigma^2 \left\{ \frac{1-b^{2n}}{1-b^2} \right\}.$$

We note immediately that for large n , $1/2 < b < 1$,

$$(9) \quad \sigma_n^2(\alpha) < \sigma_n^2(\beta),$$

so the noise level has been reduced by the subtraction.

Now, $P\alpha$ will stop about when

$$\begin{aligned}
 (10) \quad b^n u_0 &= \sigma_n(\alpha), \\
 &= \sigma \left\{ \frac{2-b^{2n}+b^{2n+1}}{1+b} \right\}^{1/2}.
 \end{aligned}$$

Neglecting b^{2n} compared to 2, and solving for n we find

$$(11) \quad n_\alpha = \frac{\log \sqrt{\frac{2\sigma^2}{(1+b)u_0^2}}}{\log b}.$$

Process $P\beta$ will stop when

$$\begin{aligned}
 (12) \quad b^n u_0 &= \sigma_n(\beta) \\
 &= \sigma \left\{ \frac{1-b^{2n}}{1-b^2} \right\}^{1/2}.
 \end{aligned}$$



UNIVERSITY OF ILLINOIS-URBANA
510.84 IL6R v.1 C002 v.100-110(1960-
Variations with size of characteristics



3 0112 088404055